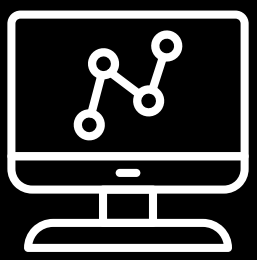
PROJET CLUSTERING R

PERROT Loick



Le clustering est une technique de datamining non supervisée qui permet de distinguer des groupes homogènes (classes, segments, clusters) au sein d'un grand volume de données.

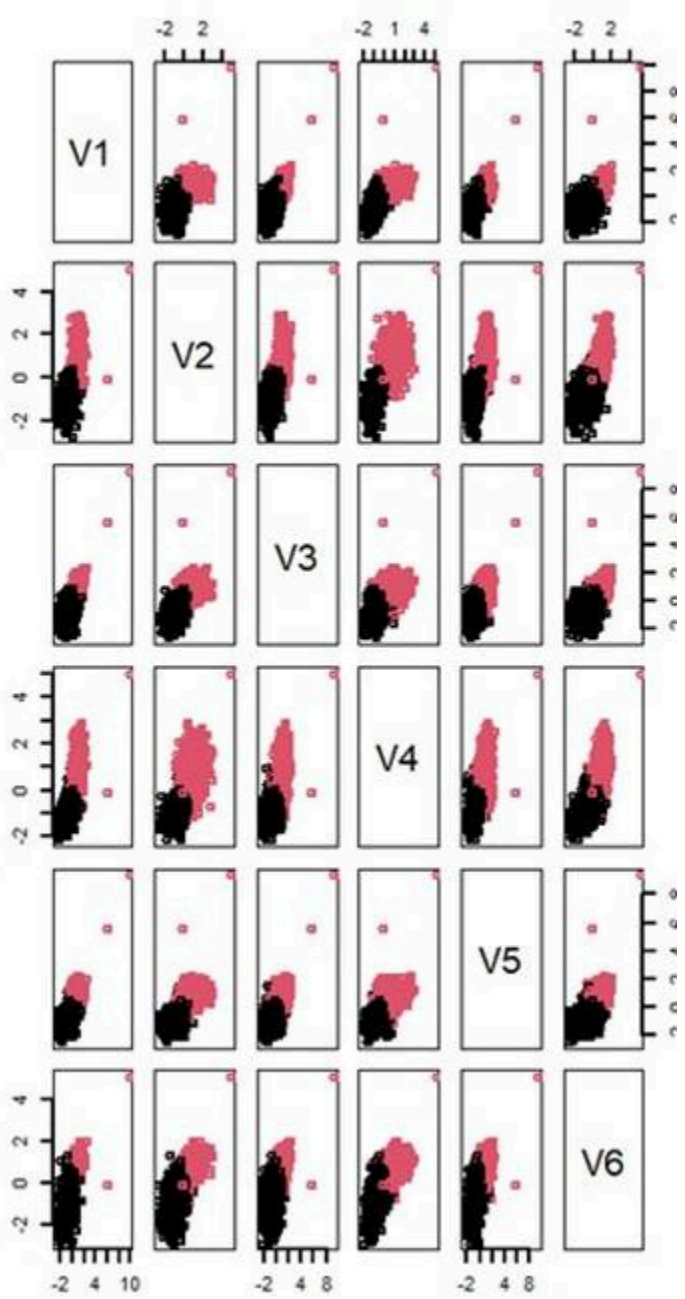
De part leur constitution, ces groupes peuvent apporter une information pertinente sur les données, notamment s'ils sont représentés graphiquement à l'aide d'une ACP.

Ils peuvent aussi servir à découper une étude en sous-parties, chacune pouvant bénéficier de traitements particuliers.

```
data=read.table("Test_Clusters_Atypiques.txt")
plot(data)
data=scale(data)
data=as.data.frame(data)
cl=kmeans(data,center = rbind(data[1,]data[1499,]),nstart=1)
cl=kmeans(data,center=2,nstart=5)
plot(data,col=cl$cluster,lwd=2)
```

Le code R ci-dessus permet de mettre en évidence deux clusters grâce à la méthode des kmeans appliquée au jeu de données "test_clusters_atypiques.txt".

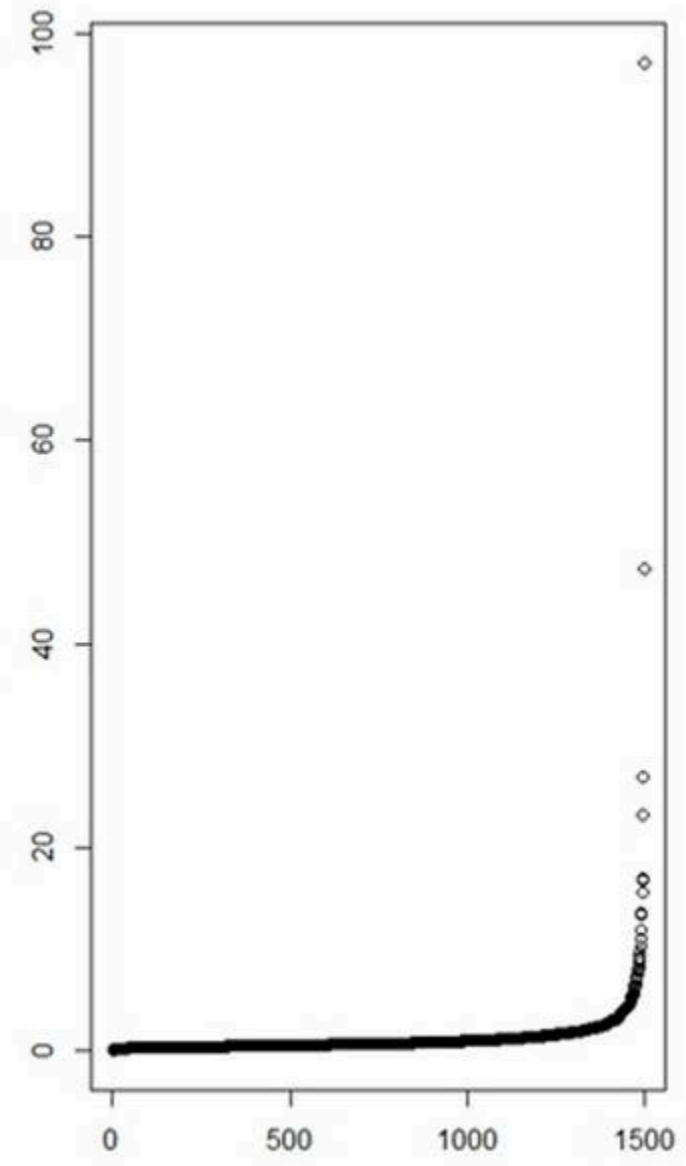
Vous pouvez observer le résultat ci-dessous.

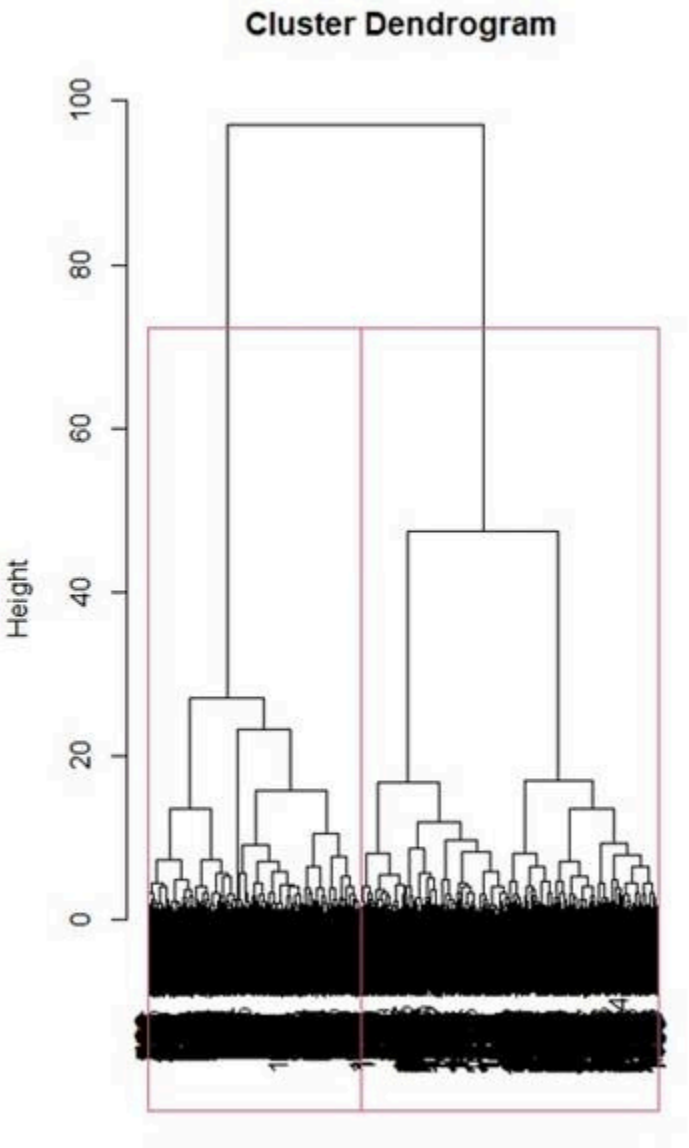

V1
V2
V3
V4
V5
V6

```r
#avec la méthode ward.D2
x=scale(data) #centre et réduit
distance=dist(x,"euclidean") #crée une structure de distance entre les individus
h=hclust(distance, "ward.D2") # crée l'arbre
plot(h$height) # affiche l'évolution du critère de dissimilarité entre classes
plot(h) # affiche le dendrogramme
rect.hclust(h,k=2) # ajoute les classes
c=cutree(h,k=2) # récupère les classes
plot(x, col = c) # affiche les points avec une couleur différente par classe indexée
# par le numéro de la classe

#avec la méthode average
x=scale(data) #centre et réduit
distance=dist(x,"euclidean") #crée une structure de distance entre les individus
h=hclust(distance, "average") # crée l'arbre
plot(h$height) # affiche l'évolution du critère de dissimilarité entre classes
plot(h) # affiche le dendrogramme
rect.hclust(h,k=2) # ajoute les classes
c=cutree(h,k=2) # récupère les classes
plot(x, col = c) # affiche les points avec une couleur différente par classe indexée
# par le numéro de la classe
```

Dans cette partie de code, nous avons utilisé l'algorithme CAH afin de mettre en évidence différents clusters. Le graphique ci-après permet de voir le nombre de cluster que contient le jeu de données.


100
80
60
40
20
0
0
500
1000
1500



Ce dernier graphique appelé dendrogramme nous permet de mettre en évidence les clusters.